ED 464 152	TM 033 837
AUTHOR	Zhang, Yanling
TITLE	DIF in a Large Scale Mathematics Assessment: The Interaction of Gender and Ethnicity.
PUB DATE	2002-04-00
NOTE	45p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Achievement Tests; Elementary Secondary Education; *Ethnicity; *Item Bias; *Mathematics Tests; *Sex Differences; Student Characteristics; Test Items
IDENTIFIERS	Delaware; *Delaware Student Testing Program; Item Bias Detection; Large Scale Assessment

ABSTRACT

A study was conducted to investigate how differential item functioning (DIF) is associated with student characteristics such as gender and ethnicity for the total group and to understand the pattern and nature of existing group differences by conducting DIF analyses separately within gender and ethnicity. The research is unique in that it undertakes a two-way DIF approach by taking into account the interaction of gender and ethnicity. Data were from the Delaware Student Testing Program. The multiple choice items for grades 3, 5, 8, and 10 were used, for a total group of 34,544 students. Results of the Mantel Haenzel DIF analyses found an approximately equal number of DIF items favoring the reference and focal groups across all four grade levels. There was some statistical evidence of DIF in all grades, but the DIF was more balanced out by gender and ethnicity in the higher grades (8 and 10). A number of gender and ethnic DIF items that were previously undetected in a total analyses were flagged when two-way procedures were applied. It is clear that two-way DIF analyses offer a more complete and comprehensive approach to DIF detection, superior to the traditional one-way DIF analysis approach and particularly useful in large-scale testing programs. (Contains 30 tables and 32 references.) (SLD)

DIF in a Large Scale Mathematics Assessment:

The Interaction of Gender and Ethnicity

Yanling Zhang

Program of Educational Research and Evaluation

Ohio University

Paper presented for the annual American Educational Research Association conference,

April 2002, New Orleans.

Objectives

The majority of recent research on mathematics performance differences, bias and/or DIF, has been focused on gender (e.g., Cole, 1997; Doolittle & Cleary, 1987; Jacklin, 1989; Linn & Hyde, 1989; Willingham & Cole, 1997), while a far smaller amount of research has been focused on ethnicity (O'Neil & McPeek, 1993; Schmitt & Dorans, 1990; etc.). Almost without exception, the studies concerning mathematics achievement found thus far on DIF have investigated gender separately from ethnicity, or vice versa. Virtually all studies conducted on DIF procedures have typically been based on aggregated gender and ethnicity data. This marginal DIF analysis ignores potential interactions between gender and ethnicity, interactions that may be important.

Analyzing DIF at the global level does not serve the purpose of illuminating actual gender and ethnic performance differences. The purpose of this observational study is twofold: (1) to investigate how DIF is associated with student characteristics such as gender and ethnicity for the total group, and (2) to better understand the pattern and nature of existing group differences by conducting DIF analyses separately within gender and ethnicity. In other words, this research is unique in that it undertakes a 2-way DIF approach by taking into account the interaction of gender and ethnicity.

Background and Theoretical Framework

To date, many states have mandated standardized achievement tests for elementary and secondary school students. Such statewide achievement tests often have high-stakes attached to their use. Ensuring a test is free from biased items is critical to the validity of the test. Test bias occurs when performance on a test requires sources of knowledge different from those intended to be measured, causing the test scores to be less valid for a particular group (Camilli & Shepard, 1994). One way to screen for bias items is to perform statistical procedures to detect differential item functioning (DIF). According to Hambleton, Swaminathan, and Rogers (1991), "an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting an item right." (p. 110).

Over the past thirty years, there has been a considerable shift in what bias means. Because of the negative connotations of the term bias, in the mid-80s, a more neutral term called differential item functioning (DIF) was proposed (Holland & Thayer, 1988), referring to items that affect performances of comparable groups differently on the trait in question. For any given large-scale assessment, DIF evaluation for a given test is suggested as a standard procedure as stated in the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985): "operational use of a test will often afford opportunities to check for group differences in test performance and to investigate whether or not these differences indicate bias" (3:10).

Among all subject areas to be assessed, mathematics is one of the most studied areas because it is an important component of fundamental education for any society. Mathematics achievement assessment also has serious consequences over students' subsequent schooling and career choices. A substantial amount of research has been conducted with respect to gender difference and/or DIF in mathematics (Carlton & Harris, 1992; Cole, 1997; Fennema & Carpenter, 1981; Jacklin, 1989; Linn & Hyde, 1989; Linn & Petersen, 1985; Maccoby & Jacklin, 1974; Sheuneman & Grima, 1997; Willingham & Cole, 1997, etc.). Previous research demonstrates that standardized achievement tests and college admissions tests given over a wide range of ages and educational levels have recognized a rough pattern of male-female difference. Overall, at the outset of schooling, very little difference has been observed between elementary girls and boys. Gender performance discrepancy begins to emerge in late elementary school.

The largest differences exist in the subject area of mathematics, where males appear more proficient than females, particularly in secondary school and beyond. In recent years, evidence has been collected, including meta-analyses, supporting the claim that such group mean difference is narrowing.

In an attempt to identify possible patterns, some researchers have tried to relate the score differences to the content of the tests (Aiken, 1986-87; Doolittle, 1989; Harris & Carlton, 1993; O'Neil & McPeek, 1993; Ryan & Fan, 1996). Findings indicate that boys are prone to experience less difficulty in geometry, and problem-solving items such as word problems and applied mathematics items whereas girls tend to perform better with conceptual, algebra questions, and items with symbols. In addition to group mean scores, emphasis has been placed on the distribution and variability of scores (Benbow, 1990; Fan, Chen, and Matsumoto, 1997; Feingold, 1992; Willingham and Cole, 1997). Research findings in this field have found that males tend to have a more spread out score distribution, with more numbers of examinees dispersed towards the higher and lower ends of the score range; females are more clustered to the center (or the mean) of the score range.

Methods

Instrumentation

The Delaware Student Testing Program (DSTP) is a mandated statewide assessment program. Data from the 1999 DSTP mathematics section were used for global and 2-way DIF analyses. The test consists of two sections: The Stanford Achievement Series, 9th edition (SAT9) of the abbreviated version in mathematical problem solving items (30 items) and Delaware developed items (MDE) (37 items). There are three item formats, MC, short answers (SA), and extended constructed response (CR) items.

The present study was set within the context of gender (male and female), ethnicity (Black, Asian, Hispanic, and White), and grade level (grades 3, 5, 8, and 10). Only the multiple choice format items were considered for analyses.

The demographic frequency distributions are listed in Tables 1 and 2 below.

Table 1

Frequency Distribution of the 1999 DSTP Examinees by Gender and Ethnicity

Grade		Gender		Total
		Male	Female	
3	Asian	96	66	162
	Black	1341	1308	2649
	Hispanic	226	219	445
	White	2835	2498	5330
	Total	4498	4091	8586
5	Asian	74	88	162
	Black	1390	1272	2662
	Hispanic	208	223	431
	White	2914	2567	5481
	Total	4586	4150	8736
8	Asian	96	78	174
	Black	1282	1283	2565
	Hispanic	167	191	358
	White	2977	2739	5716
	Total	4522	4291	8813
10	Asian	94	92	186
	Black	1216	1162	2378
	Hispanic	171	161	332
	White	2814	2601	5442
	Total	4295	4016	8338

Table 2

Frequency Distribution of Grades by Gender of the Total Group

Grade			Frequency	Percent
	Male	4504		
3	Female	4098	8,602	24.9
	Male	4594		
5	Female	4158	8,752	25.3
	Male	4529		
8	Female	4299	8,828	25.6
	Male	4337		24.2
10	Female	4025	8,362	
Total			34,544	100.0

Validity and Reliability Issues

Validity is the most important psychometric property of any measurement or assessment tool. It is worth noting that the curricula are uniform for third, fifth, and eighth grades, but not for the tenth grade. Additionally, to evaluate the construct validity of the mathematics portion of the1999 test, a principal component analysis was performed at each grade level.

The degree to which an instrument is reliable is another important indicator of the psychometric quality. The reliability was estimated for the mathematics tests of the 1999 DSTP using Cronbach's Coefficient Alpha (an index of internal consistency) at each grade level for the combined test, the SAT section and for the MDE section separately.

Statistical Analysis Procedures

Two non-IRT DIF detection statistical procedures were utilized: The Mantel-Haenszel procedure (MH) and logistic regression (LR), which are two highly recommended statistical DIF methods for MC items (Holland & Thayer, 1993; Rogers & Swaminathan, 1993).

Schmitt, Holland, and Dorans (1993) recommended that all possible examinees in each focal and reference group should be used when conducting DIF research. In this study, since the total test score is used in the DIF indices at each ability level, the largest possible number of examinees in both the reference and focal groups will be used to ensure sufficient power for stable DIF estimates.

In this study, items were treated symmetrically to identify DIF items, i.e., items having large effect in the direction of reference and focal groups were both flagged and reviewed. Afterwards, a kappa measure of agreement or decision consistency was computed to assess consistency between the two measures. Finally, an analysis of variance was performed.

The MH Procedure

Difference between the reference and focal groups may take two forms, uniform and non-uniform, that can be visually represented with item characteristic curves. When there is no interaction between ability level and group membership, uniform DIF, or ordinal DIF exists. In another word, the probability of answering an item correct consistently favor one group over the other at all levels of the conditioning variable. Non-uniform DIF, or disordinal DIF, exists when there is an interaction between ability and group membership: the probabilities of getting an answer correct vary over ability levels for the groups (Mellenburg, 1982).

The MH statistical procedure is appropriate to detect uniform DIF. It is an advantageous procedure in that not only an index of odds-ratio is provided for each reference/focal group comparison, but the index also functions as estimate of the magnitude of DIF (effect size).

The Mantel-Haenszel (MH) procedure is a contingency table method for estimating and testing a common two-factor association parameter in k 2 x 2 tables, if there are k score group levels. At each level of score-group, the reference group is assumed to be comparable to the focal group on the trait being measured by the item under consideration. An overall odds-ratio (α_{MH}) is computed for each comparison on each item. The MH chi-square is also conducted testing the null hypothesis that there is there is no DIF, Ho: $\alpha_{MH} = 1$. α_{MH} is on a scale of 0 to ∞ with α =1 meaning a null value of no DIF. A rejection of the unit α_{MH} hypothesis suggests that the DIF is present in the studied item.

Logistic Regression (LR) Procedures

MH is not sensitive to non-uniform DIF. LR is a parametric statistical approach and is powerful in detecting non-uniform DIF. In order to test non-uniform DIF, the LR procedure was used in addition to the MH.

With LR, the presence of DIF is determined by testing the improvement in model fit that occurs when a term for group membership and a term for the interaction between test score and group membership are successfully added to the logistic regression model. A chi-square test is then used to evaluate the presence of uniform and non-uniform DIF on the item of interest by successively testing each item included in the model. Performance on the studied item is first conditioned on the total test score.

The logistic regression equation used in this study is

$$Y = b_0 + b_1 \text{ TOT} + b_2 \text{ GENDER} + b_3 \text{ TOT*GENDER}$$

where Y is a natural log of the odds ratio. That is, the equation becomes

$$\text{In } [P_i / (1 - P_i)] = b_0 + b_1 \text{ TOT} + b_2 \text{ GENDER} + b_3 \text{ TOT*GENDER}$$

where p is the proportion of individuals getting the answer correct.

There are two major weaknesses in the LR DIF procedure: (a) the lack of an associated effect size measure, (b) the Type I error or false positive rates can be elevated (Jodoin & Gierl, 1999). In the context of DIF, a Type I error is the incorrect identification of an item as displaying DIF when in fact, it does not. Recently, Zumbo (1999) proposed $R^2\Delta$, a weighted least squares effect size measure for the LR DIF procedure. However, research of Jodoin and Gierl (1999) found that $R^2\Delta$ may be an unreliable measure.

The Delta Scale

Since the sensitivity of a statistical test is dependent on sample size, an effect size to distinguish statistical significance from practical significance or meaningfulness is desirable. The widely accepted delta-scale (Holland & Thayer, 1988) was adopted as the criterion for small, medium, and large DIF effect. Holland & Thayer (1988) proposed that it is convenient to take the log of α_{MH} and put it into a symmetrical scale in which zero is the null value. Thus, the transformation can be expressed as

$$\text{delta}_{MH} = -\ 2.35\ \ln(\alpha_{MH})$$

to be used as a measure of the amount of DIF. The value of estimated delta_{MH} is the average amount a member of the reference group found the studied item more difficult than did a comparable group member of focal group. A value of zero suggests no DIF is present. A negative or a positive value indicates that the test question favors the one group over another. The higher the number, the greater the difference between matched groups.

For a certain combination of reference and focal groups, all the items can be categorized into three groups.

1. Negligible DIF (A). When either delta_{MH} is not statistically significantly different from 0 or if the magnitude of the delta_{MH} values is less than 1 Δ unit in absolute value, which converts to $0.65 \leq \alpha_{MH} \leq 1.53$.
2. Medium DIF (B). All other items which are statistically significant and fall in the range of $1.53 < \alpha_{MH} < 1.89$, or $0.53 < \alpha_{MH} < 0.65$.

3. Large DIF (C). Items with delta$_{MH}$ of more than 1.5 Δ unit in absolute value, which is equivalent to $\alpha_{MH} \geq 1.89$ or $\alpha_{MH} \leq 0.53$.and is statistically significant.

For all categories, chi-square test of statistical significance is set at the 0.05 level for a single item. Because each item may have as many different DIF statistics as the number of group comparisons, the item will be categorized based on its worse case.

Matching Criterion – Thick and Thin Conditioning

In this study, the total test score was used as the matching criterion. The use of total score as the matching variable is termed as thin matching (Donoghue & Allen, 1993) whereas forming the matching variable by pooling total score levels is referred to as thick matching. Donoghue and Allen (1993) found that thin matching yielded the best results of any method examined for long tests (40 items), especially with adequate sample sizes (1600). Intermediate length tests yielded similar results for thin matching and the best methods of thick matching. For short tests (5-10 items), thin matching worked very poorly, with a tendency to falsely identify items as possessing DIF against the reference group.

The findings of Donoghue and Allen (1993) provided further support to the present study about adopting the thin matching approach in the MH approach, using total score as the conditioning variable since there were 53 MC items and sufficient sample sizes. LR naturally uses the near continuous total score as a predictor.

Purification (Inclusion or Exclusion of DIF Items)

Conditioning on ability is a critical step because it ensures that examinees are matched on a common measure before they are compared. When a large number of DIF items are found, the accuracy and appropriateness of the conditioning variable becomes questionable. After an item is shown to exhibit DIF, one strategy is to purify the criterion by removing the DIF items as they may degrade ability estimation.

Gierl, Jodoin, and Ackerman (2000) used simulated data and investigated the impact of excessive numbers of DIF items in the conditioning variable (i.e., up to 60% of the items) in terms of Type I error rates for MH, Simultaneous Item Bias Test (SIBTEST), or LR. The results showed that all three methods provide adequate Type I error protection when the proportion of DIF items is large. Therefore, purifying (removing DIF items from the conditioning variable) was not seen as a necessary preliminary procedure and was not a concern in this research.

Kappa (κ) – A Measure of Agreement

To determine the extent to which the results of MH and LR agree with each other, kappa is computed. Kappa is commonly known as Cohen's kappa (Cohen, 1960), an important statistic that is not based on chi-square, but that does use contingency tables and nominal variables.

It is defined as

$$\kappa = (P - P_c) / (1 - P_c)$$

where P represents the proportion of agreement and P_c represents the proportion of agreement due to chance. The coefficient κ is simply the proportion of agreement after adjusted by chance. It can also be expressed in frequencies to facilitate computation,

$$\kappa = (\Sigma f_O - \Sigma f_E) / (N - \Sigma f_E)$$

where f_O represents the observed frequencies on the diagonal and f_E represents the expected frequencies on the diagonal assuming that judgments are independent. In both the numerator and denominator, we subtract the same amount Σf_E, the number of agreements that we would expect merely by chance. Then a ratio of the two chance-corrected values is formed. Kappa tests H_0: κ = 0, meaning no agreement between the two variables tested.

Results

Descriptives and Variability of Scores of All Groups

Descriptive statistics can be found in the Appendix in Tables 3-6 for all grades. As a group, males had a slightly higher mean score than the females across all grades. Of all the ethnic groups and across all grades, Asians as a group had the highest mean scores, followed by the White group. Hispanics and Blacks had close group mean scores, which were noticeably lower than those of the Asian and White groups. Within each ethnic group, across all grades, males did not score consistently higher than females. The difference between males and females were not large.

In reference to variability, across all grades, males were prone to a wider variability range than females and the gap between males/females mean score standard deviations were increasing as grades go up (from 0.22, 0.60, 0.78, to 1.41). As to ethnic groups, in Grades 3 and 5, although Asians (SD=8.564) had the lowest variability than other ethnic groups, the difference between each ethnic group were not substantial at all. In Grades 8 and 10, Asians had the widest variability among all ethnic groups and the gap between the groups were much significant than lower grades. The difference between the Asian group and Black standard deviations were as big as 5.53.

Validity Estimation - The Principal Component Analysis

In order to explore the dimensionality of these items, a PCA was performed on all 53 MC items. Parallel analysis (Thompson & Daniel, 1996) was used to determine the number of factors to extract. For all grades, a distinct component with a large eigenvalue of 7.87 - 9.676, on top of a secondary component with an eigenvalue around 1.7, was extracted. In Grade 3, a component matrix showed that the majority of items loaded on the first component, with 43 out of 53 items (81%) having a correlation coefficient larger than 0.3. Although there was a second component present, only 3 items (5.66%) loaded positively onto factor 2. Similar findings echo in the results of grades 5, 8, and 10.

Reliability Estimation: The Cronbach's Alpha

The Cronbach's alpha was calculated in order to estimate the internal consistency of the mathematics test. For all grades, the SAT items had an alpha of .83 - .86 and the MDE MC items had an alpha of .75 - .79. Combining all the items as one test yielded reliability coefficients of .88 – .91 for the four grades.

Results of DIF Analyses

To make sure MH and LR produce compatible results, an actual proportion of agreement and kappa were calculated in order to estimate the agreement between the results produced by the two procedures beyond chance. Kappa was calculated to compare the MH results with the LR results of the uniform DIF. The results were summarized in Tables 7-22 in the Appendix. Across the four grades, except for the White/Asian comparisons, the actual proportion of agreement (1.00-0.830) and kappa indices (0.485-1.00) indicated a high to medium degree of agreement between the numbers of uniform DIF items flagged by both methods. Only the results of MH methods are presented here, which are shown in Tables 7-22 in the Appendix. Each table includes significant p-values ($\alpha = .05$), the corresponding effect size (Odds-ratio index α_{MH}), as well as the favoring direction of DIF.

At each grade level, five comparisons were made with respect to gender DIF, i.e., the total group gender comparison, and the gender comparison within the Asian, the Black, the

Hispanics, and the White population. Also conducted were the nine total group level ethnicity DIF comparisons and comparisons within males and females.

Grade 3.

Table 7 shows that two medium DIF (B) items were flagged as favoring males. Five large DIF (C) items were flagged, 2 favoring males and 3 favoring females. In the gender within Blacks comparison, only one B item was detected, which favored Black females. One C and one B items were flagged for the gender within Hispanics comparison, one item favoring each direction. Similar findings were found for the comparison between the White males and females.

For the White/Asian comparison (Table 8), nine B and C items were found, with four favoring Whites and five favoring Asians. The same number of B and C DIF items as well as favoring directions were found in the White males/Asian males comparison, but different items were flagged with different DIF effects. The Whites/Blacks comparison as well that within males and females only yielded zero B or C items, one B item within the males and one B item in the females. The Whites/Hispanics comparison (table 10) demonstrated two B items for the total group comparison, two B items within the males, but eight B items within the females. Among the eight B items, seven were biased against Hispanic females while one biased against White females.

Grade 5.

As shown in Table 11, zero items showed B or C DIF effects of the total male/female group comparison. Five C items were obtained for the male Asians/female Asians comparison, all favoring females. Three B items were flagged within the Black examinees only, all favoring Black females. However, Hispanic males/Hispanic females depicted a different picture. Of the

11 B and C items flagged, eight favored the Hispanic males and three favored the Hispanic females.

Table 12 lists the White/Asian comparison, and White/Asian comparison within males and females only. Three B and C items were flagged for all comparisons. Almost all of the DIF effects were occurring on different items. Two B items were found showing favor to the White total group and White males while one C item were showing favor to White females. One C item was found showing favor to Asian total group and one B item favored Asian males while one B and one C items were found showing favor to Asian females.

As shown in Table 13, no B or C items were detected for the comparisons between White/Black, White/Black within males and within females. Table 14 summarized the results of White/Hispanic comparison, White/Hispanic within males and within females comparisons. For the total White/Hispanic comparison, one C and one B items were detected, both favoring Whites. When inspecting White/Hispanic within males comparison, one C and six B items were detected, four favoring Whites and three favoring Hispanics. For the White/Hispanic within females comparison, one C and three B items were detected, 3 favoring Whites and 1 favoring Hispanics.

Grade 8.

From summary table 15, out of the male/female comparison, only two B items were detected, one in favor of each direction. Within the Asians, nine C items were distinguished, six in favor of males and three in favor of females. Among the Blacks, four B items were flagged, among which three were in favor of Black males. Eleven C and B items were flagged in the Hispanic male/female comparison, with five items in favor of males and six in favor of females. Three B items showed up among the Whites, two favoring males and 1 favoring females.

Table 16 summarized the results of 3 comparisons: White/Asian comparison, as well as that within males and within females only. Eight B items were flagged for the total group comparison, with four items in favor of each ethnic group. Five B and C items favoring of Asian males were flagged and two were in favor of White males. No B or C items were detected favoring Asian females and one C and one B were detected favoring White females.

In Table 17, it can be seen that the White/Black comparison and the White male/Black male comparison were free from any type of B or C items. Within the females, one C and one B item showed up, one favoring each ethnic group. As shown in Table 18, the White/Hispanic comparison was also free from any B or C items. Within the males, two B items exhibited preference to White males over Hispanic males. Among the females, only one B item was noticed as showing favor to the White females.

Grade 10.

For the total group male/female comparison (Table 19), none of the 53 items displayed any B or C DIF effect. When looking only within Asians, four C items were discovered, with three favoring females and one favoring males. Comparisons of male/female within the Blacks only uncovered one B item favoring females. One C and one B items, both favoring females, were found among the Hispanics. The White male/White female comparison yielded one B item favoring males.

For the White/Asian comparison in Table 20, four B and one C items were found, two favoring Whites and three favoring Asians. Within the White males and Asian males, three B and one C items were found, with two items favoring each ethnicity. Within the White female and Asian females, only two C items were found, both favoring Asians. The total White/Black comparison as well as that within males and females only, as shown in Table 124, yielded no B or C items.

In Table 22, it is shown that the White/Hispanic comparison found one B item, favoring Whites. However, within White/Hispanic males, four B items were found, three favoring Hispanic males and one favoring White males. With respect to the White/Hispanic females, three B items were uncovered, two favoring Hispanic females and one favoring White females.

Results of ANOVA

As a supplement, a two-way univariate ANOVA was conducted to investigate the main effects of gender and ethnicity, as well as gender by ethnicity interaction, on total test score, as shown in Tables 27-30. ANOVA analyses of Grade 3 revealed a statistically insignificant test of the gender by ethnicity interaction, F (8524, 3) = 1.48, p = .22. The eta squared coefficient of .001 for the gender by ethnicity interaction confirmed that there was little association between the interaction and the total score in light of the large sample size. Of the two main effects, gender and ethnicity, only ethnicity showed a significant statistical result, F (8524, 3) = 533.35, p < .001. A post hoc Bonferroni test showed that the Asian group mean was statistically higher than those of the other ethnic groups, with the largest discrepancy of 10.96 points existing between Asian and Hispanics, followed closely by a Asian and Black gap of 10.75. The difference between Asian and White group mean was 2.88. ANOVA test results of Grade 5 (Table 28) produced very similar findings to Grade 3.

ANOVA test results of Grade 8 were exhibited in Table 29. The race by gender interaction was statistically significant, F (3, 8678) = 3.87, p = .009. Partial eta squared indicates the race by gender interaction had no noticeable association by 0.1% with the total score variance. Table 30 shows the ANOVA test results of Grade 10. Similar to Grade 8, the race by gender interaction test was significant, F (3, 7915) = 7.88, $p < .000$. However, the partial eta squared showed that the factor gender by interaction only account for an insignificant amount of 0.30% of the total variances.

Summary and Conclusions

Results of the MH DIF analyses found an approximately equal number of DIF items favoring the reference and the focal groups across all four grade levels. There was some statistical evidence of DIF in all grades, where the DIF was more balanced out by gender and ethnicity in the higher grades (Grades 8 and10).

In Grades 3 and 5, two specific instances of unbalanced DIF were discovered. As shown in Table 10, Grade 3, when comparing White females with Hispanic females, seven DIF items were detected in favor of White females and one DIF item was detected in favor of Hispanic females. The group mean difference was reduced (from 7.94 to 6.55) by 17.5% after all the DIF items were deleted from the data set and the analyses were re-run. It is possible that the Hispanic females may have been disadvantaged over the White females on these items.

In Grade 5, when comparing gender within Hispanics, eight items were identified as showing favor to males and 3 items were identified as showing favor to the females (See Table 11). Prior to removing the eight DIF items, the Hispanic females had a higher mean score than the males. When the eight DIF items were removed from the total score, the female and male difference was more disparate (from 0.34 to 1.01). The White/Black comparison, and the same comparison within males and within females were the least affected by the DIF effect. This pattern was held consistent across all grades.

A number of gender and ethnic DIF items that were previously undetected in a total analysis were flagged when 2-way procedures were applied (subsequent analyses of subgroups).

Further investigation revealed that the total group gender and ethnic DIF effects were often attributed to exceptional performance by one group over the other. For instance, in Grade 3, 26 B or C items within a certain gender-by-ethnic group did not show up in the analyses of the total group. Similarly, 29 B or C items in Grade 5, 28 items in Grade 8, and 18 items in Grade 10 were flagged at the subgroup level but were not flagged at the total group level analyses. Due to the reason that the test items analyzed are still live items, the main goal of the study does not go beyond DIF identification. However, it is proposed that cases showing a disadvantage to certain ethnic groups could be reviewed by the given disadvantaged group in order to help investigate the cause of DIF.

Implications

It is clear that two-way DIF analyses offer a more complete and comprehensive approach to DIF detection, superior to the traditional one-way DIF analysis approach. The focal groups are defined by both gender and ethnicity instead of gender or ethnicity alone, so that the single unique DIF statistics associated with each focal group can be obtained. In general, two-way DIF detection procedures can benefit large-scale testing programs by detecting previously unidentified DIF items. This more complete approach to DIF analysis enhances our understanding of the nature of DIF and may offer clues as to the causes of DIF which may not be evident through one-way analyses. Furthermore, two-way DIF analysis has practical implications for both policy-makers and school educators.

Limitations of this research are primarily attributed to issues of sample size adequacy and further substantive analyses. Future studies can be continued on more nationally representative student groups so as to cross validate the findings. Understanding of the pattern and nature of

DIF items along with content analyses definitely merits further endeavors. Item formats, i.e., MC or CR items, can be included as an additional dimension.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association, Inc.

Aiken, L. (1986-1987). Sex differences in mathematical ability: A review of the literature. Educational Research Quarterly, 10, 25-35.

Benbow, C. P. (1990). Gender differences: Searching for facts. American Psychologist, 45, 988.

Camilli, G., and Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications.

Carlton, S. T., and Harris, A. M. (1992). Characteristics associated with differential item functioning on the scholastic aptitude test: Gender and majority/minority group comparisons. Princeton, NJ: Educational Testing Service.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cole, N. S. (1997). The ETS gender study: How females and males perform in educational settings. Princeton, NJ: Educational Testing Service.

Dolittle, A. E. (1989). Gender differences in performance on mathematics achievement items. Applied Measurement in Education, 2, 161-177.

Doolittle, A. E. and Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

Donoghue, J. R., and Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. Journal of Educational Statistics, 18, 131-154.

Fan, X., Chen, M., and Matsumoto, A. R. (1997). Gender differences in mathematics achievement: Findings from the National Education Longitudinal Study of 1988. The Journal of Experimental Education, 65, 229-242.

Feingold, A. (1992). Sex differences in variability in intellectual abilities. Reviews of Educational Research, 62, 61-84.

Fennema, E., and Carpenter, T. (1981). The second national assessment and sex-related differences in mathematics. Mathematics Teacher, 74, 554-559.

Gierl, M. J., Jodoin, M. G., and Ackerman, T. A. (2000). Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE Publications, Inc.

Harris, A. M., and Carlton, S. T. (1993). Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test. Applied Measurement in Education, 6(2), 137-151.

Holland, P. W., and Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), Test Validity, (pp. 129-145). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Jacklin, C. N. (1989). Female and male: issues of gender. The American Psychologist, 44 (2), 127-133.

Jodoin, M. G., and Gierl, M. J. (2000). Evaluating Type I error rates using an effect size measure with the logistic regression procedure for DIF detection. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Linn, M. C., and Petersen, A. C. (1985). Facts and assumptions about the nature of sex differences. In S. Klein (Ed.), Handbook for achieving sex equity through education (pp. 53-57). Baltimore: John Hopkins University.

Linn, M. C., and Hyde, J. S. (1989). Gender, mathematics, and science. Educational Researcher, 18(8), 17-27.

Maccoby, E. E., and Jacklin, N. C. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statstistics, 7, 105-108.

O'Neil, K. A., and McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland and H. Wainer (Eds.), Differential item functioning (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Rogers, H. J., and Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17(2), 105-116.

Ryan, K. E., and Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: a confirmatory approach. Educational Measurement: Issues and Practice, 15, 15-20, 38.

Scheuneman, J. D., and Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. Applied Measurement in Education, 10(4), 299-319.

Schmitt, A. P., and Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27, 67-81.

Schmitt, A. P., Holland, P. W., and Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland and H. Wainer (Eds.), Differential item functioning (pp. 281-315). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Thomas, B., and Daniel, L. G. (1996). Factor analytic evidence fro the construct validity of scores: A historical overview and some guidelines. Educational and Psychological Measurement, 56 (2), 197-208.

Willingham, W. W., and Cole, N. S. (1997). Gender and fair assessment. Mahwah, NJ: Erlbaum.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores [On-line]. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Available: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.

APPENDIX

Table 3

Grade 3 Group Means and Standard Deviations

Ethnicity	Gender	M	SD	N
Asian	Male	40.82	8.47	94
	Female	39.80	8.73	66
	Total	40.40	8.56	160
Black	Male	29.48	9.23	1331
	Female	29.82	9.11	1296
	Total	29.65	9.17	2627
Hispanics	Male	29.58	8.53	221
	Female	29.30	9.63	215
	Total	29.44	9.08	436
White	Male	37.77	8.79	2814
	Female	37.24	8.27	2488
	Total	37.52	8.69	5302
Total	Male	34.95	9.76	4460
	Female	34.50	9.54	4065
	Total	34.74	9.66	8525

Table 4

Grade 5 Group Means and Standard Deviations

Ethnicity	Gender	M	SD	N
Asian	Male	40.81	9.85	72
	Female	40.14	8.74	88
	Total	40.44	9.23	160
Black	Male	26.15	9.67	1375
	Female	26.83	9.18	1262
	Total	26.48	9.45	2637
Hispanics	Male	28.04	10.21	193
	Female	28.37	9.78	207
	Total	28.21	9.98	400
White	Male	35.43	9.85	3893
	Female	35.00	9.43	2554
	Total	35.23	9.65	5447
Total	Male	32.39	10.76	4533
	Female	32.27	10.16	4111
	Total	32.33	10.48	8644

Table 5

Grade 8 Group Means and Standard Deviations

Ethnicity	Gender	M	SD	N
Asian	Male	34.42	11.49	92
	Female	34.57	11.55	77
	Total	34.49	11.48	169
Black	Male	20.82	8.42	1256
	Female	21.49	7.82	1257
	Total	21.16	8.13	2513
Hispanics	Male	18.16	10.46	158
	Female	19.53	10.15	185
	Total	18.90	10.30	343
White	Male	29.25	10.28	2943
	Female	28.54	9.67	2711
	Total	28.91	10.00	5654
	Male	26.58	10.69	4449
Total	Female	26.16	9.91	4230
	Total	26.37	10.32	8679

Table 6

Grade 10 Group Means and Standard Deviations

Ethnicity	Gender	M	SD	N
Asian	Male	31.49	12.82	91
	Female	26.48	11.68	87
	Total	29.04	12.50	178
Black	Male	17.10	7.21	1117
	Female	17.71	6.72	1099
	Total	17.40	6.97	2216
Hispanics	Male	16.83	7.94	160
	Female	17.24	7.45	145
	Total	17.02	7.70	305
White	Male	24.57	10.26	2705
	Female	23.65	8.84	2512
	Total	24.13	9.62	5217
	Male	22.37	10.21	4073
Total	Female	21.77	8.80	3843
	Total	22.08	9.55	7916

Table 7

Grade 3 Male/Female Comparison, and Gender Comparisons within Each Ethnicity (DIF Favoring Direction)

Items	Male/Female	M/F within Asians	M/F within Blacks	M/F within Hispanics	M/F within Whites
	(N=4495/ 4091)	(N=96/66)	(N=1341/ 1308)	(N=226/219)	(N=2832/ 2498)
SAT2	.647* (M)	--			0.623* (M)
SAT3		5.43** (M)			--
SAT5				.560* (M)	
SAT6		10.333** (M)			--
SAT16		0.181** (F)			--
SAT19				1.905** (F)	
SAT27		0.295** (F)			--
SAT28		--			1.872* (F)
MDE21			1.603* (F)		
MDE28	.603* (M)	--			0.557* (M)
MDE29		0.402**(F)			--
Total Items Favoring Males	2	2	--	1	1
Total Items Favoring Females	--	3	1	1	2

Table 8

Grade 3 White/Asian Comparison, within Females, and within Males (DIF Favoring Direction)

Items	White/Asian (N=5330/162)	White/Asian Within Males (N=2832/96)	White/Asian within Females (N=2498/66)
SAT5	0.409** (A)	--	--
SAT6	0.455** (A)	--	0.257** (A)
SAT9	--	0.595* (A)	--
SAT17	2.167** (W)	3.174** (W)	--
SAT18	--	1.873* (W)	--
SAT19	--	--	0.350** (A)
SAT24	1.733* (W)	1.838* (W)	--
SAT27	--	0.481** (A)	--
MDE1	0.294** (A)	--	0.069** (A)
MDE3	1.735** (W)	--	--
MDE9	1.925** (W)	1.903** (W)	1.971** (W)
MDE11	1.669* (W)	1.691* (W)	--
MDE20	--	0.038** (A)	--
MDE26	0.648* (A)	0.599* (A)	--
MDE29	--	--	1.893** (W)
Total Items Favoring White	5	5	2
Total Items Favoring Asian	4	4	3

Note. All the p-values are significant at 0.05 level. Based on the delta scale, * indicates an item showing medium DIF. ** indicates an item showing large DIF.

Table 9

Grade 3 White/Black Comparison, within Males, and within Females between (DIF Favoring Direction)

Items	White/Black (N=5330/2649)	White/Black within Males (N=2832/1341)	White/Black within Females (N=2498/1308)
SAT7	--	0.585* (B)	--
SAT17	--	--	1.57* (W)
SAT22	--	1.545* (W)	--
Total Items Favoring White	0	1	1
Total Items Favoring Black	0	1	0

Table 10

Grade 3 White/Hispanic Comparison, within Males, and within Females (DIF Favoring Direction)

Items	White/Hispanic (N=5330/445)	White/Hispanic within Males (N=2832/226)	White/Hispanics within Females (N=2498/219)
SAT15	--	--	1.568* (W)
SAT17	--	--	1.576* (W)
SAT22	--	1.578* (W)	--
SAT26	--	--	0.640* (H)
SAT28	--	--	1.828* (W)
MDE2	--	--	1.617* (W)
MDE9	1.605* (W)	--	1.803* (W)
MDE20	--	--	1.839* (W)
MDE22	1.670*(W)	1.743* (W)	1.567* (W)
MDE23	--	0.644* (H)	--
Total Items Favoring White	2	2	7
Total Items Favoring Hispanics	0	1	1

Table 11

Grade 5 Male/Female, and Male/Female within Each Ethnicity (DIF Favoring Direction)

Items	Male/Female	M/F within Asians	M/F within Blacks	M/F within Hispanics	M/F within Whites
	(N=4586/ 4150)	(N=74/88)	(N=1390/1272)	(N=208/223)	(N=2914/ 2567)
SAT2				.642* (M)	
SAT4		1.894** (F)			
SAT8		3.318** (F)			
SAT11		4.904** (F)			
SAT12				.585* (M)	
SAT13				.450** (M)	
SAT14				.541* (M)	
SAT19				1.589* (F)	
SAT23			1.612* (F)		
SAT26				1.818* (F)	
SAT27			1.549* (F)		
SAT28				.493** (M)	
MDE7		2.32** (F)			
MDE8				1.718* (F)	
MDE21				.567* (M)	
MDE22				.575* (M)	.638* (M)
MDE27		5.744** (F)	1.607* (F)		
MDE28				.639* (M)	
Total Items Favoring Males	--	--	--	8	1
Total Items Favoring Females	--	5	3	3	--

Table 12

Grade 5 White/Asian Comparison, within Females and Males (DIF Favoring Direction)

Items	White v Asian (N=5481/162)	White/Asian within Males (N=2914/74)	White/Asian Within Females (N=2567/88)
SAT4	1.688* (W)		
SAT15			2.543 (W)
SAT16	.606* (A)		.449** (A)
SAT18		1.757* (W)	
SAT29			.592* (A)
MDE4	1.618* (W)	2.386* (W)	
MDE20		.360* (A)	
Total Items Favoring White	2	2	1
Total Items Favoring Asian	1	1	2

Note. All the p-values are significant at 0.05 level. Based on the delta scale, * indicates an item showing medium DIF. ** indicates an item showing large DIF.

Table 13

Grade 5 White/Black, White/Black within Males and Females (DIF Favoring Direction)

Items	White/Black (N=5481/2662)	White/Black within Males (N=2914/1390)	White/Black within Females (N=2567/1272)
Total Items Favoring White	--	--	--
Total Items Favoring Black	--	--	--

Table 14

Grade 5 White/Hispanic, within Males and Females (DIF Favoring Direction)

Items	White/Hispanic (N=5481/431)	White/Hispanic within Males (N=2914/208)	White/Hispanic within Females (N=2567/223)
SAT2		.507** (H)	
SAT4	1.705* (W)	1.612* (W)	1.759* (W)
SAT9		.634* (H)	
SAT12			1.607* (W)
SAT14		.546* (H)	
SAT19		1.702* (W)	
SAT25		1.609* (W)	
SAT28	1.970** (W)	1.670* (W)	2.308** (W)
MDE9			.644* (H)
Total Items Favoring White	2	4	3
Total Items Favoring Hispanic	--	3	1

Table 15

Grade 8 Male/Female, and Male/Female within Each Ethnicity (DIF Favoring Direction)

Items	Male/Female (N=4522/4291)	M/F within Asians (N=96/78)	M/F within Blacks (N=1282/1283)	M/F within Hispanics (N=167/191)	M/F within Whites (N=2977/2739)
SAT3				.528* (M)	
SAT6		.398** (M)			
SAT7				1.761* (F)	
SAT9		4.75** (F)			
SAT12		.318** (M)		.530** (M)	
SAT13				1.660* (F)	
SAT15		.401** (M)			
SAT16	.565* (M)				1.672* (M)
SAT18			.640* (M)		
SAT19				1.892** (F)	
SAT22		.392** (M)			
SAT24				1.84* (F)	
SAT26		.455** (M)			
SAT29	1.691* (F)		1.593* (F)	2.073** (F)	1.749* (F)
MDE2		.291** (M)	.586* (M)	.499** (M)	.538* (M)
MDE3		2.303** (F)			
MDE9			.621* (M)	.452** (M)	
MDE22		3.761** (F)			
MDE26				.566* (M)	
MDE28				1.707* (F)	
Total Items Favoring Males	1	6	3	5	2
Total Items Favoring Females	1	3	1	6	1

Table 16

Grade 8 White/Asian Comparison, within Females and Males (DIF Favoring Direction)

Items	White/Asian (N=5716/174)	White/Asian within Males (N=2977/96)	White/Asian Within Females (N=2739/78)
SAT6		.560* (A)	
SAT7	.576* (A)	.405** (A)	
SAT17	1.615* (W)		
SAT11		.600* (A)	
SAT15			1.862* (W)
SAT26			2.142** (W)
SAT29	1.658* (W)		
SAT30	1.806* (W)	1.987** (W)	
MDE2	.641* (A)	.526** (A)	
MDE20	.595* (A)		
MDE22		1.636* (W)	
MDE24	.450* (A)	.396** (A)	
MDE27	1.601* (W)		
Total Items Favoring White	4	2	2
Total Items Favoring Asian	4	5	--

Note. All the p-values are significant at 0.05 level. Based on the delta scale, * indicates an item showing medium DIF. ** indicates an item showing large DIF.

Table 17

Grade 8 White/Black, White/Black within Males and Females (DIF Favoring Direction)

Items	White/Black (N=5716/2565)	White/Black within Males (N=2977/1282)	White/Black within Females (N=2739/1283)
SAT7			.635* (B)
SAT20			1.710* (W)
Total Items Favoring White	--	--	1
Total Items Favoring Black	--	--	1

Note. All the p-values are significant at 0.05 level. Based on the delta scale, * indicates an item showing medium DIF. ** indicates an item showing large DIF.

Table 18

Grade 8 White/Hispanic, within Males and Females (DIF Favoring Direction)

Items	White/Hispanic (N=5716/358)	White/Hispanic within Males (N=2977/167)	White/Hispanic within Females (N=2739/191)
SAT3			1.704* (W)
SAT22		1.785* (W)	
MDE27		1.586* (W)	
Total Items Favoring White	--	2	1
Total Items Favoring Hispanic	--	--	--

Note. All the p-values are significant at 0.05 level. Based on the delta scale, * indicates an item showing medium DIF. ** indicates an item showing large DIF.

Table 19

Grade 10 Male/Female, and Male/Female within Each Ethnicity (DIF Favoring Direction)

Items	Male/ Female (N=4322/ 4016)	M/F within Asians (N=94/92)	M/F within Blacks (N=1216/1162)	M/F within Hispanics (N=171/161)	M/F within Whites (N=2841/2601)
SAT5		.416** (M)			
SAT7				1.791* (F)	
SAT17					.636* (M)
SAT22			1.813* (F)		
SAT30		3.714** (F)			
MDE3				1.919** (F)	
MDE6		2.965** (F)			
MDE27		2.483** (F)			
Total Items Favoring Males	--	1	--	--	1
Total Items Favoring Females	--	3	1	2	--

Table 20

Grade 10 White/Asian Comparison, within Females and Males (DIF Favoring Direction)

Items	White/Asian (N=5442/186)	White/Asian within Males (N=2841/94)	White/Asian Within Females (N=2601/92)
SAT2	.572* (A)		
SAT4	.518** (A)	.364** (A)	
SAT5		.582* (A)	
SAT11	.536* (A)		.379** (A)
SAT28	1.539* (W)		
MDE7	1.605* (W)	1.772* (W)	
MDE22		1.770* (W)	
MDE27			.526** (A)
Total Items Favoring White	2	2	--
Total Items Favoring Asian	3	2	2

Note. All the p-values are significant at 0.05 level. Based on the delta scale, * indicates an item showing medium DIF. ** indicates an item showing large DIF.

Table 21

Grade 10 White/Black, White/Black within Males and Females (DIF Favoring Direction)

Items	White/Black (N=5442/2378)	White/Black within Males (N=2841/1216)	White/Black within Females (N=2601/1162)
Total Items Favoring White	--	--	--
Total Items Favoring Black	--	--	--

Table 22

Grade 10 White/Hispanic, within Males and Females (DIF Favoring Direction)

Items	White/Hispanic (N=5442/332)	White/Hispanic within Males (N=2814/171)	White/Hispanic within Females (N=2601/161)
SAT7			.648* (H)
SAT8		.635* (H)	
SAT9		.648* (H)	
SAT28		1.653* (W)	
MDE5			.591* (H)
MDE10		.600* (H)	
MDE23			1.663* (W)
MDE28	1.714* (W)		
Total Items Favoring White	1	1	1
Total Items Favoring Hispanic	--	3	2

Note. All the p-values are significant at 0.05 level. Based on the delta scale, * indicates an item showing medium DIF. ** indicates an item showing large DIF.

Table 23

Grade 3 Kappa Analysis Results for the Agreement between MH and LR Procedures

Comparison	Proportion of Agreement	κ	p
Male/Female (M/F)	0.868	0.730	.000
M/F within Asians	0.906	0.394	.004
M/F within Blacks	1.000	1.000	.000
M/F within Hispanics	0.962	0.647	.000
M/F within Whites	1.000	1.000	.000
White/Asian	0.962	0.866	.000
White/Black	0.962	0.923	.000
White/Hispanic	0.943	0.843	.000
White/Asian within Males	1.000	1.000	.000
White/Asian within Females	0.943	0.636	.000
White/Black within Males	0.981	0.956	.000
White/Black within Females	0.906	0.792	.000
White/Hispanic within Males	0.925	0.624	.000
White/Hispanic within Females	0.887	0.677	.000

Note. All the p-values are significant at 0.05 level.

Table 24

Grade 5 Kappa Analysis Results for the Agreement between MH and LR Procedures

Comparison	Proportion of Agreement	κ	p
White/Asian	.906	.493	.000
White/Black	.962	.921	.000
White/Hispanic	.981	.912	.000
Male/Female (M/F)	.943	.883	.000
M/F within the Asians	.906	.245	.045
M/F within the Blacks	.943	.885	.000
M/F within the Hispanics	.981	.941	.000
M/F within the Whites	.925	.842	.000
White/Asian within the Males	.981	.791	.000
White/Asian within the Females	1.00	1.000	.000
White/Black within the Males	.925	.806	.000
White/Black within the Females	.925	.815	.000
White/Hispanic within the Males	.943	.790	.000
White/Hispanic within the Females	.962	.812	.000

Note. All the p-values are significant at 0.05 level.

Table 25

Grade 8 Kappa Analysis Results for the Agreement between MH and LR Procedures

Comparison	Proportion of Agreement	κ	p
Male/Female (M/F)	.981	.954	.000
White/Asian	.887	.657	.000
White/Black	.906	.810	.000
White/Hispanic	.868	.387	.005
M/F within the Asians	.943	.769	.000
M/F within the Blacks	.962	.925	.000
M/F within the Hispanics	.943	.809	.000
M/F within the Whites	.943	.881	.000
White/Asian within the Males	.943	.791	.000
White/Asian within the Females	.924	.471	.000
White/Black within the Males	.830	.635	.000
White/Black within the Females	.943	.868	.000
White/Hispanic within the Males	.943	.698	.000
White/Hispanic within the Females	.962	.485	.000

Note. All the p-values are significant at 0.05 level.

Table 26

Grade 10 Kappa Analysis Results for the Agreement between MH and LR Procedures

Comparison	Proportion of Agreement	κ	p
Male/Female (M/F)	.981	.962	.000
White/Asian	.887	.560	.000
White/Black	.924	.845	.000
White/Hispanic	.887	.563	.000
M/F within the Asians	.943	.381	.000
M/F within the Blacks	.906	.705	.000
M/F within the Hispanics	.981	.791	.000
M/F within the Whites	.962	.924	.000
White/Asian within the Males	.981	.879	.000
White/Asian within the Females	.981	.791	.000
White/Black within the Males	.924	.771	.000
White/Black within the Females	.962	.910	.000
White/Hispanic within the Males	.962	.836	.000
White/Hispanic within the Females	.924	.560	.000

Note. All the p-values are significant at 0.05 level.

Table 27

Grade 3 ANOVA Test Results

Source	SS	df	MS	F	p	Partial Eta Squared
Race	125533.322	3	41844.407	533.349	.000	.158
Sex	58.750	1	58.750	.749	.387	.000
Race*Sex	348.854	3	116.285	1.482	.217	.001
Error	668209.842	8517	78.456			
Total	795238.987	8524				

Table 28

Grade 5 ANOVA Test Results

Source	SS	df	MS	F	p	Partial Eta Squared
Race	152219.954	3	50739.985	550.921	.000	.161
Sex	.229	1	.229	.002	.960	.000
Race*Sex	583.698	3	194.566	2.113	.096	.001
Error	795378.154	8636	92.100			
Total	949418.625	8643				

Table 29

Grade 8 ANOVA Test Results

Source	SS	df	MS	F	p	Partial Eta Squared
Race	134652.250	3	44884.083	493.619	.000	.146
Gender	57.676	1	57.676	.634	.426	.000
Race*Gender	1056.146	3	352.049	3.872	.009	.001
Error	788441.966	8671	90.929			
Total	924631.628	8678				

Table 30

Grade 10 ANOVA Test Results

Source	SS	df	MS	F	p	Partial Eta Squared
Race	86100.707	3	28700.236	358.471	.000	.120
Gender	629.260	1	629.260	7.860	.005	.001
Race*Gender	1892.416	3	630.805	7.879	.000	.003
Error	633137.601	7908	80.063	7.879		
Total	722334.901	7915				

TM033837

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: DIF in a Large Scale Mathematics Assessment: The Interaction of Gender and Ethnicity

Author(s): Yanling Zhang

Corporate Source: Ohio University	Publication Date: April 1, 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: Yanling Zhang, PhD	
Organization/Address: MS: 13L, Rosedale Rd, ETS Princeton, NJ08541	Telephone: 609-683-2195	FAX: 609-683-2130
	E-Mail Address: yxzhang@ets.org	Date: 4/4/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com